基于本体的网络数据工作平台 NetData*)

范 宽 吴朝晖 陈华钧
(浙江大学计算机科学与技术学院 杭州 310027)

摘 要 近年来,网格、语义网络等新技术迅速发展并日臻成熟。互联网发展焦点开始从信息的发布和互联转向知识的交互框架。随着语义网络迅速发展,世界各地各个领域的研究爱好者组成虚拟社区,对同一领域的知识信息一起协作研究。其中,对数据的整理、保存、检索、分析是实现语义网络远景的基础工作。本文为了帮助研究社区的研究人员更有效方便地加入社区的研究,利用长期帮助中国中医研究院建设专业结构化数据库群的项目中所取得的经验,结合了语义网络和数据库网格的研究,设计并初步实现了基于本体的网络数据工作平台。

关键词 rdf, 工作台,本体,语义网,数据库网格

NetData: An Ontology-based E-Science Workbench for Semantic Web

FAN Kuan WU Zhao-Hui GHEN Hua-Jun
(Campus of Computer Science and Technology of Zhejiang University, Hangzhou 310027)

Abstract Just as the Internet is shifting its focus from information and communication to a knowledge delivery infrastructure, the Semantic Web extends the current Web to a pervasive worldwide knowledge management infrastructure which aims to allow Web entities(software agents, users and programs)to interoperate, exchange and share resources, and solve complex problems in collaborative way. The Semantic Web enables to put the experts from different domains and different locations to found cyber communities. The experts coordinate each other to do research or project, in which data manipulation (sorting, saving, retrieving and analyzing)is a fundamental building block to realize above vision. For the purpose of helping exports to join the cooperation net data work conveniently and efficiently, we design and develop an ontology-based workbench and take the advantage of experience that gained from the projects of building Traditional Chinese Medicine (TCM) Databases and the research of Semantic Web and Database Grid that have been done by our lab. Hundreds of TCM experts from different TCM Academies in China contributed to these projects through Internet.

Keywords Rdf, E-science workbench, Ontology, Semantic web, Database grid

1 简介

本体在目前的软件开发中扮演着举足轻重的角色,很多时候是整个软件架构的核心。同时,本体可以作为联结各个层次和不同应用的纽带。本文介绍的 NetData 工作台也是通过本体的导入,共享与其它平台(例如我们实验室研发的数据库网格平台)进行通讯,并结合在一起的。

NetData 工作台直接的应用需求来自中国中医研究院。中医药已经发展了几千年,在生物和医学方面为人们提供了各种各样的资源,包括各种医药材料、中医药治则、中医药方等等。大多数资源和中医药知识都以文本(包括文字和图像)的方式保存下来。在过去的几个世纪中,这些资料不停地被几代人收集、整理、扩充。现在,这些珍贵资料中的大部分已经被分布在全中国各地的中医药数字图书馆数字化并存储起来。同时,成千上万的中医院、中医药学院、研究院每天都产生相当数量的研究文章和报告。中医药数字图书馆同样将这些文章和报告作为全文保存起来。为了更好地利用这些数据资源,浙江大学 CCNT 实验室与中医研究院联合,开发了许多基于网络的、联合全国各个地区中医药专家参与的数据处理工具。实验室与中医研究院已合作研究多年,正是在开发这些工具集和多年合作所得经验的基础之上,重新整理需求并整合最新的技术,提出并初步实现了网络数据工作平台 NetData。

NetData 工作台主要由 3 个部分组成:语义支持模块、RDF 数据生成模块、存储层数据接口。

语义支持模块是比较前端的模块,它自带一个知识库(最简单的形式是词典),为用户处理数据时服务,比如自动语义标注、自动查找错误点等等。知识库的支持是与本体相关联的,比如说词典中的每一个词都与对应本体的概念和类相对应。

RDF 数据生成模块意为将用户处理过之后的输入数据生成为符合某个本体语义表达的 RDF 模型数据,产生的 RDF 数据并不直接通过存储层接口马上导入存储层(可以是传统的集中式数据库,也可以是数据库网格或者其他类型的存储层),而是保存在自带的 RDF 仓库中。

存储层接口是一个具体实现较多的模块,以适配各种不同类型的存储层。接口中比较重要的实现部分是 RDF 数据与存储层模型数据之间的转换。例如,关系型数据库的模型是关系表,首先根据 RDF 数据所依据的本体(RDFS 描述)定义该本体在关系数据模型中对应的结构(DDL),然后使用 JDBC、prepared statement 等数据库技术将相应的数据导入(SQL)。

NetData 工作台究竟能在领域专家的工作中有什么帮助呢? 在实际的项目中,该工作平台给中医药的专家提供了方便的建库环境,使在地理上分布全国各个中医药大学、研究所和图书馆的专家都参与到了中国中医研究院中医药结构化数

*)国家 863 项目(2003DEA2C015)资助课题。范 宽 硕士研究生;吴朝晖 教授、博士生导师;陈华钧 讲师、博士。

据库建库工作中来。但该系统并不与特定的专业领域挂钩，也不仅仅只能完成建库的工作。

本文就是对 NetData-Workbench 的各个方面做一整体介绍，希望借助本文，能将该系统介绍给更多的人了解、使用、改进。本文第 2 节将介绍平台所使用的本体，这些本体不仅本系统使用，实验室的其他系统也使用；第 3 节将进一步介绍系统实现的模块细节；第 4 节将介绍一些应用场景和实际应用案例；最后对本文作总结并对未来工作加以介绍和展望。

2 工作台使用的本体

语义网络作为一种文档语义的表达以及语义在 Web 应用程序以及智能代理中应用的一种途径，本体在其中扮演了显著的角色。本体已经被证明在构造和定义领域团体的元数据元语以及标准化工作中可以起到很大的作用，在跨社区搜索和整合信息方面的应用中起着关键的作用。因此，像 TCM 这样的专业领域虚拟社区也需要建立自己的共享本体[1]。

正如在简介中所说的，NetData 工作台所使用的本体同样也被其他的平台使用，比如 DartGrid。NetData 系统本身没有本体编辑器，因此使用 Protégé 2000 来建立所需的本体。Protégé 2000 使用语义网络语言 RDFS 作为本体存储的格式，但它也很容易适配和扩展到其他的语义网络语言，比如 DAML+OIL 或者 OWL。

平台需要建立的中医药本体包括中医药元语/概念、它们的定义、它们之间的关系以及中医药领域的术语和公理。这个本体包括一个抽象的中医药本体概念基本类，该基本类有 13 个最基本的属性。然后大约有 20 个左右的子类从中医药基本类继承，比如人体、心理和生理、疗法，以及中药方剂等。另外，中医药本体包含了很多特有的概念和类，比如阴阳、五行、穴位、方剂、证候等等。现在，我们已经完成整个中医药本体的类定义并且依据这些定义编辑了超过 100,000 条中医药本体实例[2]。图 1 展示了中医药本体的一部分。

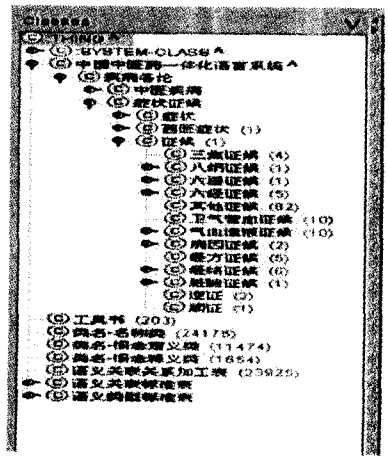


图 1 中医药本体

除了前面所说的中医药本体之外，还有一些可以看作是中医药本体的子本体。建立它们的目的是为了满足应用的需要，因为应用中有些时候需要满足应用需求的个别属性，领域的概念类大多数时候只需要囊括某一个方面的范围即可。实际上，NetData 工作台用的很多本体是在开发 DartGrid 的时候就建立了。DartGrid 是一个能够集成不同类型关系型数据库的数据库网格平台，目前主要应用在中医药领域。它依赖本体，使用 RDF 模型作为中间数据模型。需要集成的数据库将自己的数据注册到对应的本体概念，这样就组成了一个虚拟数据库[3]。图 2 展示了一个 RDFS 文件的片段。

幸运的是，这些本体也非常适合 NetData 工作台的应用。实际上，目前 NetData 在实现中就是通过 DartGrid 的本体服务来得到应用所需要的本体信息的。

```
    〈rdfs:domain rdf:resource="&acco;疾病"/〉
〈/rdf:Property〉
〈rdf:Property rdf:about="&acco;中文题名"〉
    a:maxCardinality="1"
    rdfs:label="中文题名"〉
    〈rdfs:domain rdf:resource="&acco;文献"/〉
    〈/rdf:Property〉
〈rdfs:Class rdf:about="&acco;中药"〉
    rdfs:label="中药"〉
    〈rdfs:subClassOf rdf:resource="&tcm; DartClass"/〉
〈/rdfs:Class〉
〈rdf:Property rdf:about="&acco;中药 ID"
    a:maxCardinality="1"
    rdfs:label="中药 ID"〉
    〈rdfs:range rdf:resource="&v2;ValuNode"/〉
    〈rdfs:domain rdf:resource="&acco;中药"/〉
〈/rdfs:Property〉
〈rdf:Property rdf:about="&acco;中药不良反太应"
    a:maxCardinality="1"
    rdfs:label="中药不良反应"〉
    〈rdfs:range rdf:resource="&v2; VelueNode/"〉
    〈rdfs:domain rdf:rdsource="&acco;中药"/〉
〈/rdf:Property〉
```

图 2 RDFS 文件片断

3 工作台系统架构

如在第 1 节简介中所述，NetData 工作台总体上可以划分为 3 个模块：前端的语义支持，RDF 数据生成、存储层接口。在更具体的设计划分和实际实现中，3 个模块共享一些组件，它们互有交叉。本节就对系统的设计、结构等方面做进一步的介绍。图 3 展示了系统的整个架构，可以看出本体在其中的关键作用。

3.1 语义支持

语义支持在本系统中可以分为 3 个层次。首先，系统可以是完全没有语义支持的，语义的正确性等完全由系统的使用者控制；其次是可以根据本体保证数据之间关系和结构的完整和正确，并在出现错误的情况下可以自动报警；最后是在 KB 的支持下采用算法，尽可能地辅助数据加工。例如，利用自动语义标注预处理全文文本。所以，这个模块的组成部分就包括了本体、知识库(标准词和同义词词典)和语义引擎。

语义引擎是语义支持的核心模块，它结合用户的输入、对应的本体信息以及 KB 提供的资源来分析语义问题或者产生辅助的语义信息。知识库包括标准词和同义词库等，是与领域相关的，应该由领域专家来提供，其本身就是语义性很强的数据。本体，准确地说应该是与系统本体服务通讯的接口，可以看成是知识库的必要组成部分。语义支持是在网络环境下的应用，又因为与用户的交互较复杂，所以借鉴了 KIM 的客户端形式，采用 IE 的插件[6]。

从以上的描述可以发现，语义支持模块是一个可以自我生长的子系统。换句话说，因为系统的语义支持是分层次的，从完全没有语义支持到有算法参与的语义辅助加工，所以可以利用系统本身来加工系统所需的语义数据，包括标准词、同义词和规则库等等。

3.2 RDF 数据生成

RDF 数据生成由以下这些组件组成：本体、RDF 解释表达器、RDF 生成器、RDF 检索器和 RDF 仓库。

当然，RDF 数据生成的核心就是 RDF 数据生成器，它需

要知道数据所依据的本体信息，因此也需要与本体打交道。RDF 数据生成器依据用户输入数据所对应的本体信息，将用户输入的数据转换成 RDF 模型的数据，同时将生成的 RDF 数据保存在系统自带的 RDF 仓库中，而不是将数据直接导入到存储层。

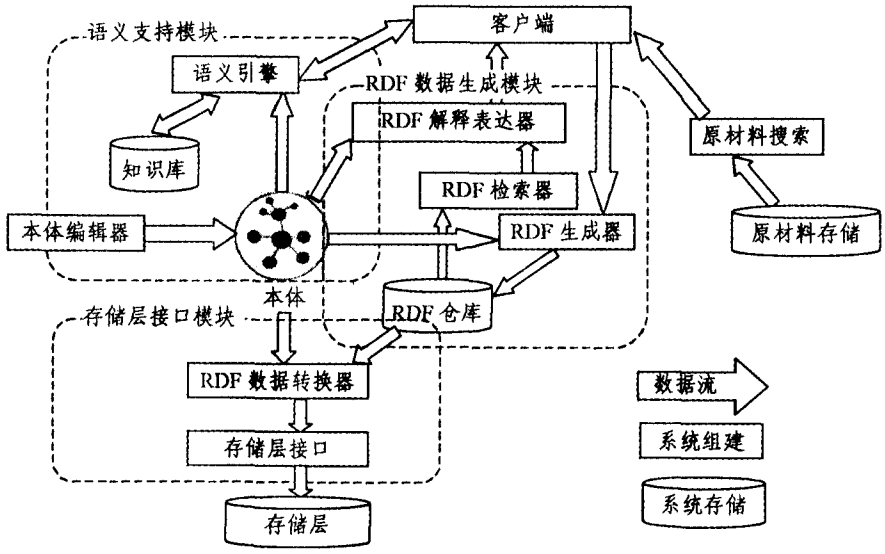


图 3　工作台系统架构

这样做是出于两个方面的考虑：一是数据质量，数据处理之后需要其他专家的审校才能确保数据质量，因此生成的 RDF 数据有质量等级，必须要有缓存，RDF 仓库就是扮演缓存的角色；二是数据安全，生成的 RDF 数据保存在单独的仓库，可以作为数据备份，提高数据的安全性，RDF 仓库也扮演备份数据的角色。仓库中的 RDF 数据被标记，例如数据质量等级、处理人员、处理日期等，所以与 RDF 仓库相关还有一个简单的 RDF 数据查询器，主要通过标记的数据进行查询。

RDF 的解释表达器主要是为了两个目的而存在：一个是为了领域专家在处理数据之后能对结果有一个直观的了解，另一个是为了方便其他专家查看、审校数据。因为中医药的应用是 Web 应用，所以我们为领域专家提供 Web 用户界面。

RDF 数据生成模块需要能够与存储接口模块通讯，否则工作台的意义就将大打折扣，因为工作台目前的应用就在于使用语义网的一些技术维护更新传统的存储层。RDF 数据生成的输入是领域专家对原始材料处理之后的输出；RDF 数据生成的输出成为存储接口的输入。

3.3　存储层接口

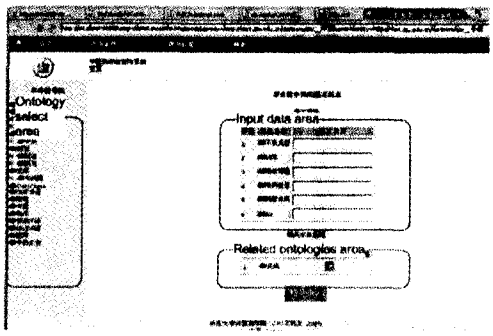


图 4　实现的用户界面

存储层接口最主要的功能就是将 RDF 模型数据转换为存储层模型的数据，然后将转换后的数据导入到存储层合适的位置。所以存储层接口包括了两个组件：RDF 数据转换器和存储层适配器。

RDF 数据转换器结合本体信息和存储层适配器提供的目标格式和数据模型信息，将 RDF 数据转换为目标数据，然后存储层适配器再确定数据要导入的位置以及如何导入数据到存储层。

4　实现和应用

实验室实现了工作台的基础版本，并在这个版本之上建立了中医药的应用。基础版本实现了前面提到的系统的 3 个模块。实现基础版本使用了 java 的相关技术以及一些开源的项目平台，比如 spring framework，jena 等等。主要的开发工具是 Eclipse3. 1IED。

jena 是一个建设语义网的工具，主要作为 RDF 引擎来处理 RDF 模型数据。RDF 数据生成模块就是在 jena 的基础之上开发的。RDF 数据仓库使用 mySQL 来存储管理。

就像在本文简介中提到的，存储层接口模块需要多种实现来适配不同类型的存储层。工作台本身除了 RDF 仓库之外并没有其它的存储组件。在基础版本的实现中，实现了对两种存储层的适配：一种是关系型数据库，另一种就是实验室开发的数据库网格 DartGrid。现以 DartGrid 为例说明存储层导入数据的步骤。首先，存储层适配器询问 DartGrid 的数据库索引服务，得到目标数据库的 URL 和类型。然后，RDF 数据转换器得到这些信息并结合本体信息来转换目标数据。最后，存储层将数据以及相关信息提交给 DartGrid 的数据更新服务。

语义支持模块目前基本实现了第二个层次，也就是利用标准词和同义词库检查中医药语法错误。

目前在工作台上开发的应用是帮助中医药领域专家建设中医药数据库。因为是 Web 应用，系统使用 Tomcat 作为 Web 容器。图 4 展示了中医药应用的 Web 界面。当然，只要有合适的领域本体，工作台就可以应用到其它的领域。另外，除了建设数据库，工作台还可以进行其他数据处理的工作，例如数据评估。

总结和未来的工作 本文介绍了 NetData-Workbench，即一个语义网环境的网上数据工作台。设计和开发这个平台的目的是为了帮助领域专家更加方面快捷地在线加入领域研究和相关工程。文内介绍了工作台的体系和工作台使用的本体。并且利用这个平台建立了中医药领域的应用。本体是平台的核心部分，没有本体，工作台就不能成为一个整体。

当然，还有很多的工作需要继续完成。语义支持是很重要的部分，也是平台可以不断发展的部分，进一步的开发要不断地加强这一模块的功能。另一方面，在下一步的开发中打算使用 OWL 替代 RDFS，作为本体的描述语言来得到更强的语义表达。同时，我们对于数据的表达形式要做到可配置，使得领域专家在处理数据时可以更加方便有效。

所以，接下去最需要改进平台的两个部分：一个是语义支持，另一个就是数据的表现形式和数据的输入方式需要可以对应于本体概念进行配置。

参 考 文 献

1 Chen Huajun, Wu Zhaohui, Huang Chang, et al. TCM-Grid: Weaving a Medical Grid for Traditional Chinese Medicine. In: International Conference on Computational Science, 2003. 1143～1152
2 Wu Zhaohui, Chen Huajun, Xu Jiefeng. Knowledge Base Grid: A Generic Grid Architecture for Semantic Web. J Comput Sci & Technol, 2003, 18(4): 462～473
3 Wu Zhaohui, Chen Huajun, Deng Shuiguang, et al. DartGrid: RDF-Mediated Database Integration and Process Coordination Using Grid as the Platform. In: APWeb 2005. 351～363
4 Dingli A, Ciravegna F, Wilks Y. Automatic semantic annotation using unsupervised information extraction and integration. In: Proceedings of SemAnnot 2003 Workshop, 2003
5 Decker S, Melnik S, van Harmelen Frank, et al. The Semantic Web: The Roles of XML and RDF. IEEE Internet Computing, 2000, 4(5): 63～74
6 Kiryakov A, Popov B, Ognyanoff D, et al. Semantic Annotation, Indexing, and Retrieval. International Semantic Web Conference (ISWC), 2003

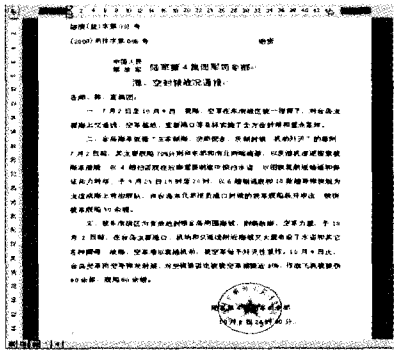
(上接第 84 页)

CryptSignMessage(签名原始数据)

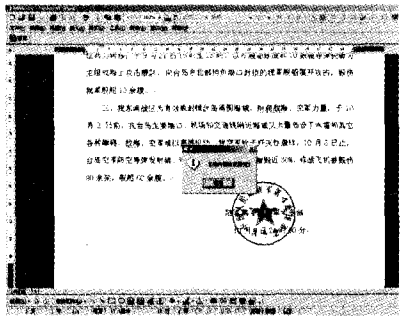
CryptVerifyMessageSignature(对签名进行验证)

3. 实现结果

为验证该电子签章的实际效果，笔者利用该电子签章对军用文书分别进行签章、验证和篡改后的验证，实验结果如图 5 所示。



(a) 文档的签名



(b) 文档内容无篡改的验证

(c) 文档内容篡改后的验证

图 5 WORD 文档的签名与验证

4 基于 PKI 的电子签章的安全性分析

基于 PKI 技术的电子签章的安全性主要表现在以下几个方面：

1. 数字证书和用户印章本身的安全。本电子签章中，用户证书和印章图像绑定在一起，采用 USB 接口的硬件设备－电子令牌作为存储介质。数字证书及其对应的私钥、用户印章图像都存储在电子令牌中，私钥由电子令牌内置 CPU 生成，不可导出数据，使得电子令牌无法被复制及仿冒；同时，电子令牌中数据由用户可以定期修改的 PIN 码提供保护，可有效地防止电子令牌丢失后被他人冒用。

2. 信息的完整性。由 Hash 函数(散列函数)的特性可知，若信息在传输过程中被篡改，完整性受到破坏，接收方重新计算出的摘要必然不同于用发送方的公钥解密出的摘要，则接收方得知其得到的信息并非发送方最初发送的信息。

3. 信源确认(信息可认证性)。因为公钥和私钥间存在对应关系，既然接收方能用发送方的公钥解开加密的摘要，并且其值与接收方重新计算出的摘要一致，则该信息必然是发送方发出的。

4. 信息的不可抵赖性。由于只有信息的发送方持有自己的私钥，其它人不能冒用其身份，故发送方无法否认他曾经发送过该信息。

结束语 本文设计并实现了一个基于 PKI 技术的电子签章系统，该系统将数字证书与印章图像绑定在一起，保证了签名的不可抵赖性；同时电子签章又确保了文档的完整性，使得对签章以后的文档的任何改动都会从签章上显示出来。在网络安全服务倍受关注的今天，该系统具有较高实用价值，有着广阔的应用前景。

参 考 文 献

1 关振胜. 公钥基础设施 PKI 与认证机构 CA [M]. 北京：电子工业出版社，2002
2 张世永. 网络安全原理与应用[M]. 北京：科学出版社，2003
3 雪涛. 基于 PKI 的安全电子邮件系统的设计与实现[D]. 四川：四川大学计算机系，2003
4 曹丽红. 浅议《电子签名法》[J]. 网络安全技术与应用，2004，46(10)：69～70
5 周城，郭正荣. 基于 Microsoft 密码体系的数字信封的实现[J]. 重庆大学学报，2005，28(6)：77～78